Learning by Example: Training Users with High-quality Query Suggestions

Morgan Harvey*
MIS Department
Northumbria University
Newcastle, UK
pwhq2@northumbria.ac.uk

Claudia Hauff
Web Information Systems
TU Delft
The Netherlands
c.hauff@tudelft.nl

David Elsweiler
I:IMSK University of
Regensburg
Germany
david@elsweiler.co.uk

ABSTRACT

The queries submitted by users to search engines often poorly describe their information needs and represent a potential bottleneck in the system. In this paper we investigate to what extent it is possible to aid users in learning how to formulate better queries by providing examples of high-quality queries interactively during a number of search sessions. By means of several controlled user studies we collect quantitative and qualitative evidence that shows: (1) study participants are able to identify and abstract qualities of queries that make them highly effective, (2) after seeing high-quality example queries participants are able to themselves create queries that are highly effective, and, (3) those queries look similar to expert queries as defined in the literature. We conclude by discussing what the findings mean in the context of the design of interactive search systems.

Categories and Subject Descriptors: H.3.3 Information Search and Retrieval
General Terms: Measurement, Experimentation, Human Factors
Keywords: Search expertise; Reflection; Behavioural Change; User Study

1. INTRODUCTION

Much of the IR research in the last half century has, with great success, focused on developing improved retrieval models to enhance the utility of retrieval systems for the end user [41]. In this line of research search queries submitted to a retrieval system are considered as a given. The focus is placed on what to do systematically to return relevant documents given this limited representation of the user's information need. A complementary approach with potentially more scope for future performance gains is to focus on giving the system more to work with by assisting users in creating better queries for specific search systems [32, 24].

Considerable evidence exists showing that many users do not know how to generate good queries. Analyses of search transaction logs show that people use short queries, especially on the Web [2] and even in this familiar domain a good proportion of searches fail completely [13]. In many search domains, including Web and Email search (with domain-specific search systems and interfaces), expert users achieve better retrieval effectiveness than novices and demonstrate different querying behaviour [3, 12, 43]. Moreover, despite the fact that most users today have to navigate through a range of search systems in their digital life, it has been reported that many users are inflexible in their approach and tend to use the same querying strategies regardless of task and available search system [29].

Typical solutions to assist users in creating effective search queries are the use of search UI features, such as query suggestions [35], related searches [36] or query autocompletion [5]. Alternatively, systems can employ context and personalisation techniques [11], which involve storing (and learning from) personal search histories and preferences to understand what a user knows and likes [18].

A third approach is to educate users about how to become better searchers [28] or to help users reflect on their own behaviour by comparing it to experts [6]. This method has the advantage that it is complementary to technical solutions. Our research continues along this path by investigating to what extent we can teach users how to pose better search queries to a particular search system. In contrast to existing approaches, we aim to understand if users are able to recognise, compare and contrast the properties of their own queries with good queries (provided by the system) and make changes to the queries they generate based on these insights. This is a new way of thinking about query suggestions; instead of providing automated examples for users to simply click on, we present them in a way that leads the user to reflect on his own behaviour, positively influencing his actions as a result.

The two principle research questions we answer are:

RQ1 Are users able to notice differences between good queries and their own and abstract these differences to change their own behaviour? If so, what are the noticeable differences?

RQ2 How effectively can users learn and abstract from good queries; do users who are "trained" perform better than users who did not receive the training? Which properties of their queries do users adapt after training?

*All three authors contributed equally to this work.

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.
 ISBN 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767731.

2. RELATED WORK

It is well-recognised that searchers have difficulties communicating their information needs [7, 38, 24]. Taylor writes of a series of stages a user goes through when seeking information. These range from experiencing a visceral need, which is "probably inexpressible in linguistic terms" to a compromised need - a "representation of the inquirer's need within the constraints of the system and its files" [38]. Therefore, in order to generate successful queries, the user must overcome several cognitive challenges: 1) to determine himself what the need is and what kind of document will solve it; 2) to choose terms that describe that document well out of a very large set of possibilities [15] and 3) to communicate using the system's vocabulary and not his own [9].

Many interactive solutions have been designed to help the user overcome these challenges and improve the representations of information needs systems have to work with. The following subsection briefly reviews such work.

2.1 Interactive Query Support

IR systems can attain better descriptions of information needs by explicitly asking for certain details. The I3R system offered a means for users to provide terms and concepts they felt were important and identify relationships between these terms and other concepts in the domain [10]. Similarly, Kelly and Fu [24] used clarification forms to elicit additional information about the search context from users. The forms queried users on what they knew and what they would like to know about the topic and why. These were shown to be helpful in achieving improved retrieval performance.

A second technique is to assist the user to iteratively improve their own queries by adding additional terms suggested by the system, commonly referred to as interactive query expansion (IQE) [17]. This approach gives the user much more control over the search than if the query were to be expanded automatically (i.e. where the system selects expansion terms without user input [31]). Although IQE can offer improved performance [25], it has been shown that users are poor at identifying the terms that will offer the best improvement to their queries [33, 1]. This finding is intriguing with respect to our aims as it begs the question of whether or not users are able to identify qualities of good terms or whether they just assume terms suggested by a system will automatically be of a high quality.

Relevance feedback systems [34] are a further means to expand queries without explicitly choosing terms. Instead, relevance judgements are solicited on the returned documents. In addition to expanding queries, other scholars have investigated the performance of systems suggesting similar or related queries e.g. [36].

Improving user queries need not be achieved via technical solutions. One group in the 2007 SIGIR workshop breakout group identified a spectrum of possible solutions from manually-led approaches (based on improved information literacy and teaching) through to automatic, system-based approaches (based on more intelligent systems) [32]. The following section reviews literature on changing user behaviour via primarily non-technical means.

2.2 Changing Behaviour

Behaviour change support systems are "information systems designed to form, alter, or reinforce attitudes or behaviours or both without using coercion or deception" [30]. Within the context of search, changes can be made to the underlying retrieval engine or to the interface to 'nudge' people towards submitting longer or better queries or to look deeper in the results list [6]. Altering the size [14] and wording [8] of the search box, for example, has been shown to influence the length of queries submitted. Moreover providing a simple "Google-like" search interface as opposed to a complicated multi-field catalogue search can radically alter user behaviour [27]. Training users on how to construct queries can improve search behaviour [26]. For example, providing guidance on the advanced features that can help with specific search tasks can improve performance for these tasks and users are able to preserve and use the knowledge gained weeks later [28]. Moreover, allowing users to reflect on their own behaviour and, importantly, compare their behaviour to other, expert users, enables individuals to improve their own habits. In [6] users, after reflection, spent longer considering search results and issued longer queries. They also used a wider range of techniques and search engine features.

We extend some of the ideas in [6] here. Rather than inviting users to compare their behaviour with that of experts, however, we investigate if they are able to learn by comparing their own queries to examples generated by the system to be near optimal for the task at hand. In doing so we relate the kinds of approaches shown in Section 2.1 with the approaches in this section. We attempt to 'nudge' users to improve their queries via high-quality examples shown via widgets similar to those described above.

3. RESEARCH METHODOLOGY

The aim of our work is to establish whether showing users of an *unfamiliar* search system examples of high-quality queries (for a small number of information needs) enables them to create better-performing queries themselves. We investigate to what extent users learn more successful querying behaviours from those examples.

Based on our two research questions (Section 1) we devised the following research hypotheses:

H1 Users are able to adapt their querying behaviour to pose good queries to an unfamiliar search system.

H2 Users are able to identify characteristics of high-performing queries that allow them to perform so well.

H3 A small number of "training queries" is sufficient to enable a user to learn how to pose good queries themselves.

H4 A user who receives training with queries he can relate to (i.e. that are anticipated to perform well), learns *better* than a user receiving training with queries that are not predicted to perform well.

H5 A user who receives training with queries he can relate to, learns *faster* than a user receiving training with queries that are not predicted to perform well.

We conducted a number of user studies (Figure 1), each requiring the *automatic generation of high-quality queries* for a given information need and search system (described in Section 4.1). To address the issue of predicted performance, we performed an initial user study (Section 4.2) to investigate participants' perceptions of the generated queries. In contrast to the later studies, participants were not given access to our search system, their judgement was solely based on their own past experience.

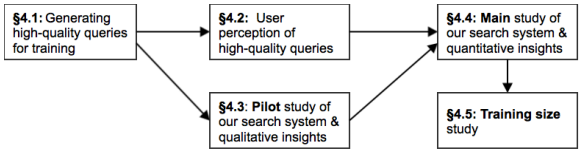

Figure 1: Overview of our experimental design.

Concurrently with the *User Perception Study*, we performed a *Pilot Study* (Section 4.3) which gave us qualitative insights into the characteristics of good queries that users were able to identify. The results of these two studies then allowed us to conduct a larger *Main Study* (Section 4.4) and a follow-up *Variable Training Size Study* (Section 4.5) with a consistent design, but different training parameters. The aim here was to better understand how much training is required to achieve an effect. In each of these studies participants were asked to perform a series of ad-hoc retrieval tasks using our search system.

To maintain maximum control over the experiments and have access to complete statistics of the collection the participants were searching over, we used a standard test collection: AQUAINT[1] together with the 50 TREC 2005 Robust track queries [40]. As our indexing and search engine we chose Apache SOLR[2]. To provide our study participants with a familiar user interface for searching the collection, we developed a web-based front-end in PHP (Figure 3).

4. EXPERIMENTS & RESULTS

In this section, we present an overview of each study and its results in turn.

4.1 Generating High-Performing Queries

In generating the "high-performing" query examples, we make the assumption that a query q_a is better than another query q_b for a given information need if q_a returns a higher Average Precision (AP) score. It is also important that the queries are understandable by humans and are not excessively long. Therefore, we are not interested in queries that happen to return good results because of a statistical anomaly or because they are overly verbose and specific.

Candidate queries were obtained via a recursive, greedy search algorithm. For each topic and its corresponding set of relevant documents, a collection was built consisting of only those relevant documents. The query building process is initiated by first considering only queries of length 1 (i.e. single-term queries) and choosing each of the top 100 terms from the topic-specific document collection (after stop words had been removed). Each initialisation of the recursive method takes in a base query and adds each of the top 100 terms to it. All 100 new potential queries are run against the entire collection using the standard SOLR search system and the AP score of the top 50 returned documents is computed. The list of queries is then ranked by their AP values and the top 10 are added to the candidate query list. Subsequently, the algorithm is initiated again with new base query having the newly-selected term added to the end. This recursive process was continued up to a query length of 4. At the end of the process any duplicate queries were removed and the top 100 queries (according to AP) were selected as the final list of candidates.

Note that this approach differs significantly from previous methods proposed in the literature for generating queries, e.g. [4], as our goal is fundamentally different. Rather than generating queries which appear to be samples from the collection (i.e. stochastically drawn from collection statistics), we are specifically interested in queries which yield high performance, are understandable and would, potentially, be posed by real users. Other related approaches used to find optimal queries in Boolean systems (e.g. [37]) were inappropriate due to differences in the underlying retrieval system. While our greedy approach does not produce globally optimal queries, it quickly produces large numbers of queries with an AP score of around 0.4. Concrete examples of generated queries can be found in the last column of Table 4.

Considering the top 100 queries for each topic, the median AP obtained by the generated queries over the first 20 returned results was 0.389. On a per-topic basis, the median was 0.391, the lowest average achieved was 0.054 and the maximum was 0.948 (IQR=0.31). Overall, 28 out of 50 topics had at least one query with an AP score greater than 0.5 and only 11 topics had any queries in the top 100 with an AP score below 0.2.

4.2 User Perception of Queries

To gain insights into how users perceive our high-quality queries (and as a precursor to answering hypotheses **H4** and **H5**), we conducted a crowd-sourcing experiment on the CrowdFlower[3] platform.

4.2.1 Study Overview

Each crowd-sourced task consisted of one search topic (in natural language form) and one of the queries generated in Section 4.1. Specifically, the workers were instructed as follows:

> *You already know query suggestions from search engines such as Google that present you with suggested queries while you type or show related queries alongside search results.*
>
> *In this task, you will be given an information need (in natural language form) and a query suggestion that has been derived for this information need. You are asked to judge the query suggestion along three dimensions - surprise, usage and relevance.*

Four questions had to be answered on a 5-point Likert scale:

1. How much do you know about the topic of the information need? (1: Very little, 5: A lot)
2. How surprised are you about the generated query suggestion? (1: Not at all surprised, 5: Extremely surprised)
3. Would you use this suggestion in an actual search? (1: No, I would not use it, 5: Yes I would use it)
4. What do you think the search result quality will be if this suggestion is used as query? (1: Very low quality, 5: Very high quality)

[1]We removed duplicate documents in a pre-processing step, to provide a better and more familiar user experience.
[2]`http://lucene.apache.org/solr/`
[3]`http://www.crowdflower.com`

Each job consisted of 10 tasks and workers were paid 12 cents (a standard rate). In this and all following CrowdFlower experiments the participants were restricted to countries where English is a native language.

For each of the Robust track topics, the 15 most effective queries generated were judged by CrowdFlower workers. Each query was judged by 3 workers, and thus, for each topic 45 judgements were collected. Three examples of topics, generated suggestions and worker ratings are shown in Table 4.

4.2.2 Results

Our workers found many of the search topics rather challenging with an average topic knowledge rating of 2.21. The most familiar topics tended to be of broad interest to many different communities; the two with the highest average knowledge ratings (3.00 and 2.89 respectively) were *What factors contributed to the growth of consumer on-line shopping?* (topic 639) and *Identify drugs used in the treatment of mental illness.* (topic 383). In contrast, search topics focusing on very specific themes or entities tended to elicit the lowest familiarity ratings; the topic with the lowest average knowledge rating (1.58) was *What is the status of The Three Gorges Project?* (topic 416).

When considering how unexpected the presented suggestions were (i.e. the "surprise" factor) we found that the vast majority of queries (more than 80%) were at least somewhat expected, receiving a rating between 1 and 3 (top-left of Figure 2). Only a small number of suggestions were considered to be *extremely surprising* and those were mostly found in topics our study participants knew little about. This indicates that our query generation approach is achieving its goal of generating queries understandable to humans.

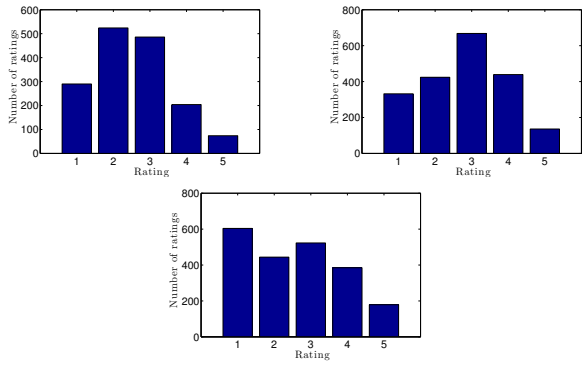

Figure 2: Histograms of "surprise" (top-left), "search quality" (top-right) and "suggestion usage" (bottom) ratings across the 750 different generated query suggestions, each rated by three users.

Of note is that fewer than 7% of judgements estimated the queries to achieve a *very high quality* of search results (top-right of Figure 2), while in contrast nearly 17% of the judgements were rated as likely to return *very low quality* search results. This result indicates that users are not able to judge the quality of query suggestions well, corroborating previous findings that users are unable to differentiate good search terms from bad ones [33, 1]. This result can only be partially explained by their lack of topical domain knowledge as the correlation between knowledge ratings and search quality ratings was moderate (but significant) with $r = 0.35$.

Lastly, we consider the question of to what extent users would use the shown suggestions in an actual search (bottom of Figure 2). Not unexpectedly, the correlation between the estimated search result quality and the potential usage is high ($r = 0.77$). Based on the ratings we have to conclude that many suggestions are not convincing, only a small number would definitely be used (9% of those rated 5) while 30% would definitely not be used (ratings of 1).

In summary, we find that user perception of our high-quality queries varies; many of them are not recognised as being effective. We make use of this result in the *Main Study*: one group of users receives high-quality suggestions recognised as high quality in this study, while another group of users receives high-quality suggestions that were rated as low quality in this study.

4.3 Pilot Study

The pilot study had three goals: (i) to test the validity of our system and task setup, (ii) to learn more about experimental factors such as participant fatigue, and (iii) most importantly, to collect qualitative data in order to establish whether participants are able to notice qualities of example queries that make them so effective as hypothesised in **H2**.

4.3.1 Study overview

The participants (n=22) consisted of university students and staff members recruited via email lists and announcements in lectures from a major European university. Although the participants were not native English speakers, all had advanced English language skills. They were given access to our search system and asked to complete 10 search tasks. As seen in Figure 3 the information need was prominently displayed to the participants. Each time they issued a query *(1)*, its retrieval effectiveness was displayed *(5)* in terms of the number of returned relevant documents within the top 20 results and the average precision (which was referred to as "search performance score"). Any relevant documents returned by the search were highlighted in blue *(4)*.

The participants were instructed to submit queries that they believed would return relevant documents (i.e. useful and containing information pertinent to the task). They were told that the documents had already been evaluated for relevance and that each submitted query would be scored in terms of how many relevant documents were returned in the top 20 results and the positions of those documents within the ranked list. This second score is simply average precision as used during the automatic query generation process and users were encouraged to focus on this to determine how well they were doing in the task. Users could move on to the next topic with a click on the *New topic, please* button *(6)*. Due to the interactive nature of the study, we selected 10 of the 50 Robust TREC topics by first eliminating those that that were either very difficult or very easy for our search system (measured in average precision achieved when using the title of the search topic as query) and then drawing randomly from the remaining topics[4].

The study participants were provided with query suggestions as shown in Figure 3 *(3)* similarly to how Web search

[4]The topics used were 303*, 362, 367*, 375*, 378, 383*, 401, 426*, 638* and 689. * indicates those that were later also used in the *Main Study*.

Figure 3: Screenshot of search interface showing a list of search results as well as some query suggestions.

engines often present query suggestions. After participants have posed their first two queries to the system for a particular topic, they are shown a number of our high-quality query suggestions. All displayed suggestions are *more effective* (achieve an AP at least 10% higher) than the participant's previous queries. Thus, different participants receive different suggestions, depending on the quality of their initial queries. The interface conveys to the participant that these are high-quality queries and they are encouraged to use them (Figure 3 *(3)*).

To test hypothesis **H2**, i.e. to establish whether users are able to learn from high quality query examples, after every use of a suggestion participants were prompted to describe in a text box *why* they considered it to be effective: "*You used the suggested query* [query]. *Considering your previous queries for this topic (shown below), what do you think is it about the suggested query that makes it so effective?*".

4.3.2 Results

The pilot findings help fine-tune our setup for the *Main Study*. Overall, the setup worked well, however we did establish fatigue to be a considerable factor. Figure 4 plots for each topic in sequence (recall that study participants receive the 10 topics in random order) the AP achieved by all queries submitted for the n^{th} topic across all study participants. It is evident that over time (i.e. queries submitted for later topics) the retrieval effectiveness degrades. In particular after the 7^{th} topic, the median AP is close to zero.

To investigate hypothesis **H2** we analysed the free-text explanations from participants describing why they believe the example queries performed so well. The responses show that participants were indeed able to identify positive query characteristics. In total 81 descriptions were supplied and out of the 22 participants, 15 gave at least one description of a suggestion. 3 participants gave descriptions for all of the suggested queries they used.

We analysed the responses qualitatively using an affinity diagramming technique, a process allowing the discovery and validation of patterns in qualitative data [16]. 12 codes were generated describing qualities participants assigned to high-performing suggestions. These are shown in Table 1.

Category
C1: Specific query terms (specification)
C2: More general query terms (generalisation)
C3: Queries not in topic description
C4: Unexpected or surprising vocabulary
C5: Surprising non-use of vocabulary
C6: Uses term the user was surprised at the usefulness of (i.e. perhaps not surprising given the topic, but surprising that it was good for performance)
C7: Thinking creatively
C8: Advanced vocabulary (rare but not on a specialist subject relating to the topic)
C9: Specialist vocabulary (rare and to do with a specialist subject relating to the topic)
C10: Good combination of search terms
C11: Using synonyms and related concepts
C12: Query requires specialist or background knowledge

Table 1: Overview of the query categories identified during the pilot study

Not only does the established coding scheme provide evidence that users *are* capable of noticing and abstracting differences between the suggested queries and their own - a prerequisite to learning - but the responses given are similar to those reported in the literature as being useful query reformulation strategies [23] or typical for queries submitted

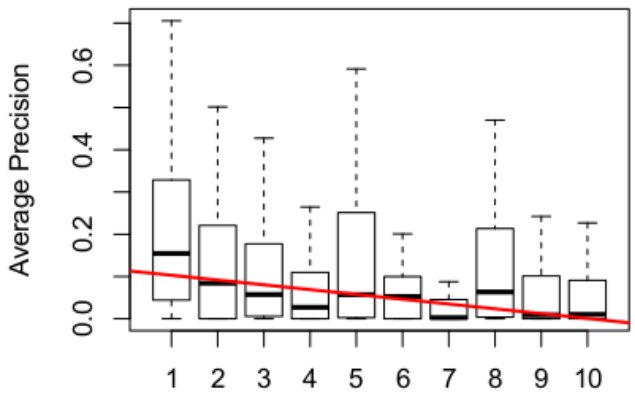

Figure 4: Pilot study: Average precision over sequences of topics showing fatigue. The nth element of the box plot contains the AP achieved over all queries across all users submitted for the nth topic the users worked on (since topics were issued in randomised order, the topic sequence differs per user).

by system and domain experts. For example, a common way to improve queries is to either make them more specfic (C1) or general (C2) [23]. Experts submit queries which are more elaborate [21] (C7, C11, C12), use broader and more varied vocabulary [39] (C1, C2), exploit synonyms and related concepts [22](C11), and include terms not used in topic descriptions [21]. Moreover, domain experts often search with queries containing specialist or domain knowledge [42] (C9, C12).

We take this as evidence to accept hypothesis **H2**. It is important to point out, however, that some of the participants explicitly mentioned in their responses that they would not be able to create some of the examples due to lack of domain knowledge or vocabulary (C10, C13).

We conclude that, despite the fact that participants are not universally able to recognise good queries (Section 4.2.2), our pilot data show that for many queries people can determine a range of properties that explain good performance.

4.4 Main Study

The main study addresses hypotheses **H1**, **H3**, **H4** & **H5** and draws from the outcomes of the two previously discussed user studies.

4.4.1 Study overview

In this study, we use search topics that our workers in the user perception study considered themselves knowledgeable about to reduce the potential influence of domain knowledge on our results. We base our choice of experimental conditions on the reported perceptions of queries to reflect **H2** and we reduced the number of tasks to six in an effort to counteract the fatigue effect observed in the pilot.

We use a between-groups design with participants randomly assigned to one of three experimental conditions:

- Group G_{exp_high}: this experimental group receives high-quality query suggestions in the training phase which were predicted to be effective in the user perceptions study (Section 4.2).
- Group G_{exp_low}: this experimental group receives high-quality query suggestions in the training phase which were predicted to be ineffective in the user perceptions study (Section 4.2).
- Group $G_{control}$: the control group does not receive any query suggestions.

For groups G_{exp_high} and G_{exp_low}, where suggestions are given, we split tasks into two phases: the first four topics are considered the *training phase*, where suggestions are shown, and the final two tasks are referred to as the *test phase*, where no suggestions are presented. Suggestions are provided using the same approach and interface as in the *Pilot Study*, i.e. suggestions were only given after two freely-created queries had been submitted and when there were queries available that would increase the AP score by at least 10%. Again, topics were issued in random order.

The participants (n=91, 29 in G_{exp_high}, 34 in G_{exp_low} and 28 in $G_{control}$) were also recruited via CrowdFlower and were paid 50 cents for the completion of a job. A job consisted of using our search system on six ad-hoc retrieval tasks; the study participants were not informed about the two phases of the study, they simply performed six search tasks (after four of which the query suggestion UI element was removed).

4.4.2 Results

We first compare the effectiveness of the issued queries, before looking at properties of the submitted queries and the fatigue factor.

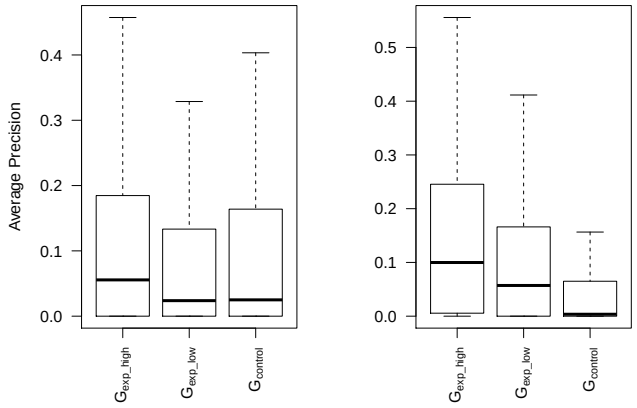

Figure 5: Main study: Querying performance over groups. Left: training topics. Right: test topics.

Effectiveness of Submitted Queries.

The fairest way to compare the performance across groups is to consider only the first 2 queries submitted by each participant for each topic. Doing so means we only consider queries submitted before suggestions are provided for a topic. Kruskal-Wallace rank sum tests show no significant difference between the groups on the training topics (p-value=0.320) but a significant difference for the test topics (p-value=0.002), with both experimental groups (G_{exp_high} and G_{exp_low}) performing significantly better than $G_{control}$.

If we consider all queries submitted for the test topics, not just the first 2 (as now no suggestions are shown to any user group), then these results become even clearer as shown in

the top half of Table 2; participants of the G_{exp_high} group issue on average queries achieving an AP of 0.10, while participants of the alternative experimental condition G_{exp_low} achieve an AP of 0.06. The control group $G_{control}$ at this stage submits queries which are an order of a magnitude worse, with a mean AP of 0.004.

Figure 5 presents an alternative view of the submitted queries' effectiveness across groups; the left boxplot shows the retrieval effectiveness for the training topics whereas on the right the effectiveness for the testing topics is shown. It is evident that participants who receive high-quality training suggestions perform better on average, but also that they are able to achieve much higher maximum average precision scores.

Main study	**Training 2 Queries**	**Testing 2 Queries**	**Testing All Queries**
1: G_{exp_high}	0.0560	0.043	0.0998†
2: G_{exp_low}	0.0238	0.041	0.0571†
3: $G_{control}$	0.025	0.024	0.0039
Training-size study	**Training 2 Queries**	**Testing 2 Queries**	**Testing All Queries**
1: G_{exp_high}	0.0922	0.055†	0.0591†
2: G_{exp_low}	0.014	0.0343	0.0666†
3: $G_{control}$	0.04	0.0132	0.0132

Table 2: Average AP values aggregated across the first two queries of the training topics (column II), the first two queries of the test topics (column III) and all queries submitted for the test topics (column IV). † indicates a significant improvement over the $G_{control}$ condition (Kruskal-Wallace rank sum test, p-value $\ll$ 0.01).

If we look at how retrieval effectiveness changes as participants query more on the same topic, we see a strong trend where G_{exp_high} and G_{exp_low} continue to improve while those in $G_{control}$ do not (Figure 6). At query position 1 there is very little difference between the groups; $G_{control}$ is only scoring on average 0.005 worse than G_{exp_high}. However, this pattern changes quickly with the experimental groups able to achieve steadily more effective queries the more they submit, which is not the case for the control group $G_{control}$. By the 4^{th} query the difference between $G_{exp_high}$and $G_{control}$ widens considerably to 0.135.

These findings provide strong evidence of retrieval effectiveness improvements for the experimental groups over the control group. The analyses so far, however, do not evidence a significant difference in performance gain between experimental conditions G_{exp_high} and G_{exp_low}.

Properties of Submitted Queries.

Beyond simply considering the retrieval effectiveness attained by a given query, we can also look at other properties of it that relate to its effectiveness or quality. These properties (shown in Table 3) go some way towards explaining the observed improvements in performance achieved by both experimental groups. We evaluated the submitted queries with metrics reflecting the literature on expert querying behaviour (see Section 4.3.2). On many of these metrics the experimental groups G_{exp_high} and G_{exp_low} significantly outperform the control group $G_{control}$. The trend is generally that group G_{exp_high} scores highest, group G_{exp_low} scores slightly

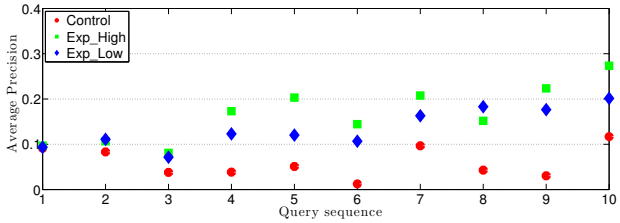

Figure 6: Main study: Average precision over sequences of queries on test topics. Each point in the plot represents the mean AP of all queries submitted as nth query. Truncated at query 10 as later queries have very few data points associated with them.

lower, but often not significantly so, and group $G_{control}$ achieves the poorest scores. Participants in G_{exp_high} and G_{exp_low}, for example, tended to submit longer queries (in both words and characters), which is noteworthy as the example queries they were shown were designed not to be overly long.

Out of all three groups, participants in G_{exp_high} submitted the rarest query terms. We measured this both in terms of the IDF statistics for the collection (i.e. their query terms feature significantly less often in the test corpus as a whole) and in terms of the number of overall participants who submitted those terms (we refer to this as median *UserCount-Term* in Table 3). Comparing the Jaccard-coefficient scores for query and topic description terms across the experimental conditions reveals that participants of G_{exp_high} were also the most likely to take terms from the topic descriptions given to them. These results suggest that a good query creation strategy was to use rare terms and seek inspiration from the topic descriptions, echoing findings from the literature [44]. While this could be negatively construed, since topic descriptions do not exist in real-life and users actually have difficulties in describing what they want [38], this finding does not explain the whole picture as there is no significant correlation (r=0.21) between AP and the overlap of queries with the topic descriptions (Jaccard score).

G_{exp_high} participants also submitted significantly more queries per topic than $G_{control}$ participants. However, this is less likely to explain the performance gains as there is no significant difference in the median number of queries submitted between G_{exp_high} and G_{exp_low} nor between G_{exp_low} and $G_{control}$. From the median time per topics it is also evident that G_{exp_high} and G_{exp_low} spend significantly more time working on each topic than $G_{control}$.

Fatigue.

One factor that could potentially affect the results is that of fatigue; are groups G_{exp_high} and G_{exp_low} doing better because they are feeling less fatigued by the task, perhaps as a result of getting some assistance in the early topics or by being shown that high performance was possible and thus increasing motivation? There are a number of metrics we can consider to try to ascertain if fatigue is present: the amount of time spent per query (query duration) and the amount of time spent per topic (topic duration) and the number of queries submitted. For all 3 groups the median query duration does seem to decrease slightly over the topics - linear models show a significant negative coefficient over the top-

	G_1: G_{exp_high}	G_2: G_{exp_low}	G_3: $G_{control}$	Wilcox (p-value) G_1/G_2	G_1/G_3	G_2/G_3
Median query length (words)	5.4	4.4	4.3	$p < 0.05$	$p < 0.01$	—
Median query length (chars)	29	28	23	—	—	$p < 0.05$
Median #queries per topic	3, IQR=4	2, IQR=4	2, IQR=2	—	$p < 0.05$	—
Median time per topic (seconds)	165	150	97	$p < 0.01$	$p < 0.01$	$p < 0.01$
Median time per query (seconds)	13	11	13	$p < 0.01$	—	—
Median query term IDFs	4.9	5.24	5.15	$p < 0.01$	—	—
Median UserCountTerm	43	45	51	—	$p < 0.01$	—
Jaccard coefficient (query terms, topic descr. terms)	0.3	0.25	0.25	$p < 0.01$	$p < 0.01$	—

Table 3: Main study: Overview of query properties aggregated for each user group across the two topics issued during the test phase.

ics of between -0.6 and -0.99. This is not the case for topic duration, however, as there is no significant trend for any of the 3 groups meaning that they all spend roughly the same amount of time on each topic. The same consistency is also present when looking at the number of queries submitted. There is no significant correlation between topic sequence and number of queries for any of the groups although groups G_{exp_high} and G_{exp_low} do submit more queries overall. These factors do not point strongly to fatigue being a factor, although the subtle changes in query duration do suggest that users are spending less time thinking about each query as time goes on, which may explain the consistent reduction in average precision.

4.5 Variable Training Size Study

An obvious question to ask, given these results, is what impact does the number of training topics given to the test groups have on performance. A final study investigated to what extent the number of training topics (hypothesis **H3**) influences a user's ability to formulate good queries.

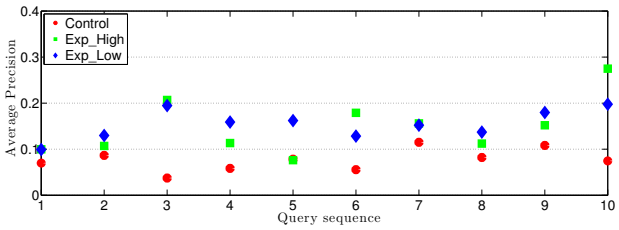

Figure 7: Training-size study: Average precision over sequences of queries on test topics. Each point in the plot represents the mean AP of all queries submitted as nth query. Truncated at query 10 as later queries have very few data points associated with them.

4.5.1 Study overview

We used the same setup and experimental design as in the *Main Study* and varied only the ratio between training and test topics: in this study we used two topics for training, and the remaining four topics for testing. As in the *Main Study*, participants (n=57, 19 participants in each condition) were recruited via CrowdFlower.

4.5.2 Results

The results from this study were analysed in the same fashion as those from the main study as can be seen in the bottom half of Table 2. The major finding of the *Main Study* holds in this experiment as well: both experimental groups outperform the control group wrt. effectiveness. Thus, even a very small amount of training (2 topics) is useful and aids users in learning to formulate better queries.

In contrast to the *Main Study*, and unsurprising given the lower amount of training, we observe a smaller difference in retrieval effectiveness across the test topics: 0.05 (G_{exp_high}), 0.054 (G_{exp_low}) and 0.024 ($G_{control}$) respectively.

These results suggest that some form of learning is taking place and that the relative improvements are smaller if less training is given. They also serve to further highlight the unexpected finding that there is little difference between G_{exp_high} and G_{exp_low}.

5. SUMMARY AND DISCUSSION

Our findings vs. our research hypotheses.

Hypothesis **H1** has been shown to hold - users are indeed able to adapt their search behaviour to an unfamiliar search system. While $G_{control}$ (which received no training) does not adapt, we clearly see significant changes in querying behaviour in both experimental groups (i.e. those who received training).

Our pilot study served to confirm hypothesis **H2**; the study participants were indeed able to determine a set of characteristics that well-performing queries contain. Recognising such characteristics is a necessary requirement for learning how to create better queries in general and not just for specific topics.

The main study and the follow-up focusing on the training set size provide evidence for hypothesis **H3**. The two experimental groups outperform $G_{control}$ significantly, both when being shown two and four training topics respectively. Thus, even a very small set of training topics is sufficient to improve users' ability to pose good queries.

Our results do not support **H4**. In terms of AP, although Figure 5 hints that G_{exp_high} may have outperformed G_{exp_low},

ID	Information need	av. KNOW	av. SUR	av. QUAL	Query suggestion examples
303	*Identify positive accomplishments of the Hubble telescope since it was launched in 1991*	2.64	2.62	2.98	[*universe astronomer faint hubble*], [*infrared galaxies universe hubble*], [*infrared stars universe hubble*]
383	*Identify drugs used in the treatment of mental illness.*	2.89	2.45	3.36	[*antidepressant risk zoloft prozac*], [*zoloft studies prozac*], [*antidepressant effective zoloft prozac*]
416	*What is the status of The Three Gorges Project?*	1.58	3.09	2.60	[*cofferdams damming generating 2009*], [*dam corporation phase 2009*], [*2009 river construction*]

Table 4: Examples of Robust track search tasks and the generated high-quality query suggestions. Columns 3-5 contain user rating data from our study on user perceptions of queries. Column 3 (KNOW) contains the average knowledge rating of the information need across all users of the study. Columns 4 and 5 contain the average rating users assigned to all query suggestions of the topic with respect to the surprise (SUR) factor and the estimated result quality (QUAL).

the difference is not significant. There were some features of the queries that were statistically distinguishable between these groups, but we feel that the evidence is not strong enough to claim that **H4** holds.

Finally, based on the evidence in Figures 6 and 7, we have to reject **H5** - our participants in both experimental groups had a comparable learning rate (though with different absolute performance scores).

Our findings vs. prior work.

Previous work has presented mixed evidence for people's ability to accurately determine which query terms will have utility. Our findings suggest this is a complex behaviour. Although participants were able to identify positive characteristics of queries shown to be effective (Section 4.3.2), many high-performing queries were not predicted to be such (Section 4.2.2). Perhaps these potentially contradictory findings indicate a potential systems bias, i.e. do users implicitly trust suggestions presented by the system as good? Is it only when doubt is introduced by explicitly questioning users about the queries that they perceive suggestions to be potentially not of good quality? What does this mean for the learning effect? This line of thought opens up many fascinating questions of how query suggestions are presented.

Our work has added to the small base of literature demonstrating means for users to learn how to provide higher-quality queries. One limitation of our work has to do with the time period of learning. Our findings support the claim that being shown good suggestions can lead to users learning how to produce better queries, however this is only demonstrated over the period of a session i.e. the test groups achieved better performance for later queries and for later topics. Ideally, however, what we want to show is learning over longer periods of time, such as weeks [28] and months [6] as previous studies have done. This requires a different mode of evaluation as crowd-sourcing is not suited to such tasks and represents an important next stage in our project.

A further limitation, with respect to how our findings may be used, is that in a real-life scenario a search system would normally not have access to relevance judgements. This means our method of creating queries cannot typically be applied. We argue that there are situations, though, that may be ideally suited to such an approach. For example in web search we have implicit indicators for difficult tasks (i.e. where better queries might be required) [20] and we also have good models for determining search success based on user behaviour [19]. When such instances combine (i.e. when users are successful in tasks they have been struggling with), this might be the perfect time to present a query suggestion, perhaps along the line of "The following query would get this page further up the ranking". Another potential use-case might be to present examples when a user switches context or to a new search-engine where new strategies are required. It has been suggested that users tend not to vary their strategy [29] and our approach might help encourage more diverse or tailored behaviour.

6. CONCLUSIONS

The set of user studies described in this paper have demonstrated that it is possible to use high-quality query examples to influence the queries users submit themselves. We have shown that users can recognise and abstract positive qualities of good queries. Users change the properties of the queries they submit and achieve better retrieval performance after seeing good examples for other tasks. Our findings open up a range of interesting questions relating to how query examples should be presented and how this affects learning and the influence of learning duration, i.e. is user behaviour influenced over the longer term? Finally is domain knowledge an important factor? We hope to address these issues in upcoming work.

7. REFERENCES

[1] P. Anick, *Using terminological feedback for web search refinement: a log-based study*, SIGIR '03, ACM, 2003, pp. 88–95.

[2] A. Arampatzis and J. Kamps, *A study of query length*, SIGIR '08, ACM, 2008, pp. 811–812.

[3] A. Aula, N. Jhaveri, and M. Käki, *Information search and re-access strategies of experienced web users*, WWW '05, ACM, 2005, pp. 583–592.

[4] L. Azzopardi, M. De Rijke, and K. Balog, *Building simulated queries for known-item topics: an analysis using six european languages*, SIGIR '07, ACM, 2007, pp. 455–462.

[5] H. Bast and I. Weber, *Type less, find more: fast autocompletion search with a succinct index*, SIGIR '06, ACM, 2006, pp. 364–371.

[6] S. Bateman, J. Teevan, and R.W. White, *The search dashboard: how reflection and comparison impact search behavior*, SIGCHI, ACM, 2012, pp. 1785–1794.

[7] N.J. Belkin, *Helping people find what they don't know*, Commun. ACM **43** (2000), no. 8, 58–61.

[8] N.J. Belkin, D. Kelly, G Kim, J-Y Kim, H-J Lee, G. Muresan, M-C Tang, X-J Yuan, and C. Cool, *Query length in interactive information retrieval*, SIGIR '03, ACM, 2003, pp. 205–212.

[9] J.L. Bennett, *The user interface in interactive systems*, Annual review of information science and technology **7** (1972), no. 159-196.

[10] W.B. Croft and R.H. Thompson, *I3r: A new approach to the design of document retrieval systems*, JASIST **38** (1987), no. 6, 389–404.

[11] Z. Dou, R. Song, and J. Wen, *A large-scale evaluation and analysis of personalized search strategies*, WWW '07, ACM, 2007, pp. 581–590.

[12] D. Elsweiler, *Supporting human memory in personal information management*, Ph.D. thesis, University of Strathclyde, 2007.

[13] B.M. Evans and E.H. Chi, *An elaborated model of social search*, IP&M **46** (2010), no. 6, 656–678.

[14] K. Franzen and J. Karlgren, *Verbosity and interface design*, SICS Research Report (2000).

[15] G.W. Furnas, T.K. Landauer, L.M. Gomez, and S.T. Dumais, *The vocabulary problem in human-system communication*, Communications of the ACM **30** (1987), no. 11, 964–971.

[16] J. Hackos and J. Redish, *User and task analysis for interface design*, John Wiley & Sons, Inc., 1998.

[17] D. Harman, *Towards interactive query expansion*, SIGIR '88, ACM, 1988, pp. 321–331.

[18] M. Harvey, F. Crestani, and M.J. Carman, *Building user profiles from topic models for personalised search*, CIKM '13, ACM, 2013, pp. 2309–2314.

[19] A. Hassan, R. Jones, and K. L. Klinkner, *Beyond dcg: user behavior as a predictor of a successful search*, WSDM, ACM, 2010, pp. 221–230.

[20] A. Hassan, R. W. White, S. T Dumais, and Y. Wang, *Struggling or exploring?: disambiguating long search sessions*, WSDM, ACM, 2014, pp. 53–62.

[21] H.A. Hembrooke, L.A. Granka, G.K. Gay, and E.D. Liddy, *The effects of expertise and feedback on search term selection and subsequent learning*, JASIST **56** (2005), no. 8, 861–871.

[22] I. Hsieh-yee, *Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers*, JASIST **44** (1993), 161–174.

[23] B.J. Jansen, D.L. Booth, and A. Spink, *Patterns of query reformulation during web searching*, Journal of the American Society for Information Science and Technology **60** (2009), no. 7, 1358–1371.

[24] D. Kelly and X. Fu, *Eliciting better information need descriptions from users of information search systems*, IP&M **43** (2007), no. 1, 30–46.

[25] J. Koenemann and N.J. Belkin, *A case for interaction: a study of interactive information retrieval behavior and effectiveness*, SIGCHI '96, ACM, 1996, pp. 205–212.

[26] W. Lucas and H. Topi, *Training for web search: Will it get you in shape?*, JASIST **55** (2004), no. 13, 1183–1198.

[27] D. McKay and G. Buchanan, *Boxing clever: how searchers use and adapt to a one-box library search*, OzCHIÕ13, ACM, 2013, pp. 497–506.

[28] N. Moraveji, D. Russell, J. Bien, and D. Mease, *Measuring improvement in user search performance resulting from optimal search tips*, SIGIR '11, ACM, 2011, pp. 355–364.

[29] J. Nielsen, *Incompetent research skills curb users' problem solving http://www.useit.com/alertbox/search-skills.html last accessed january, 2015*, Alertbox (2011).

[30] H. Oinas-Kukkonen and M. Harjumaa, *Towards deeper understanding of persuasion in software and information systems*, ACHI '08, IEEE, 2008, pp. 200–205.

[31] S.E. Robertson, *On term selection for query expansion*, Journal of Documentation **46** (1990), no. 4, 359–364.

[32] K. Rodden, I. Ruthven, and R.W. White, *Workshop on web information seeking and interaction*, ACM SIGIR Forum, vol. 41, ACM, 2007, pp. 63–67.

[33] I. Ruthven, *Re-examining the potential effectiveness of interactive query expansion.*

[34] I. Ruthven and M. Lalmas, *A survey on the use of relevance feedback for information access systems*, The Knowledge Engineering Review **18** (2003), no. 02, 95–145.

[35] B. Shneiderman, *Dynamic queries for visual information seeking*, Software, IEEE **11** (1994), no. 6, 70–77.

[36] B. Smyth, E. Balfe, J. Freyne, P. Briggs, M. Coyle, and O. Boydell, *Exploiting query repetition and regularity in an adaptive community-based web search engine*, User Modeling and User-Adapted Interaction **14** (2004), no. 5, 383–423.

[37] E. Sormunen, *A method for measuring wide range performance of boolean queries in full-text databases*, Tampere University Press, 2000.

[38] R.S. Taylor, *Question-negotiation and information seeking in libraries*, College & Research Libraries **29** (1968), no. 3, 178–194.

[39] P. Vakkari, *Changes in search tactics and relevance judgements when preparing a research proposal a summary of the findings of a longitudinal study*, Information Retrieval **4** (2001), no. 3-4, 295–310.

[40] E.M. Voorhees, *The trec 2005 robust track*, SIGIR Forum **40** (2006), no. 1, 41–48.

[41] E.M. Voorhees, D.K. Harman, et al., *Trec: Experiment and evaluation in information retrieval*, vol. 63, MIT press Cambridge, 2005.

[42] R.W. White, S.T. Dumais, and J. Teevan, *Characterizing the influence of domain expertise on web search behavior*, WSDM '09, ACM, 2009, pp. 132–141.

[43] R.W. White and D. Morris, *Investigating the querying and browsing behavior of advanced search engine users*, SIGIR '07, ACM, 2007, pp. 255–262.

[44] P. Willett and I. Ruthven, *Relevance behaviour in trec*, Journal of Documentation **70** (2014), no. 6, 1098–1117.